Assignment 6: Generalized Linear Models

Rebecca Marx

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics (ENV872L) on generalized linear models.

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Use the lesson as a guide. It contains code that can be modified to complete the assignment.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Be sure to **answer the questions** in this assignment document. Space for your answers is provided in this document and is indicated by the ">" character. If you need a second paragraph be sure to start the first line with ">". You should notice that the answer is highlighted in green by RStudio.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file. You will need to have the correct software installed to do this (see Software Installation Guide) Press the `Knit` button in the RStudio scripting panel. This will save the PDF output in your Assignments folder.
6. After Knitting, please submit the completed exercise (PDF file) to the dropbox in Sakai. Please add your last name into the file name (e.g., "Salk_A06_GLMs.pdf") prior to submission.

The completed exercise is due on Tuesday, 26 February, 2019 before class begins.

Set up your session

1. Set up your session. Upload the EPA Ecotox dataset for Neonicotinoids and the NTL-LTER raw data file for chemistry/physics.

2. Build a ggplot theme and set it as your default theme.

```
#1
setwd("C:/Users/rsmar/OneDrive/Documents/Spring 2019/RFolder/Environmental_Data_Analytics")
library(RColorBrewer)
library(viridis)
```

```
## Loading required package: viridisLite
```

```
library(colormap)
library(gridExtra)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------- tidyverse_conflicts() --
## x dplyr::combine() masks gridExtra::combine()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

```
#install.packages("dunn.test")
library(dunn.test)
```

```r
#get data
EPAecotox <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Mortality_raw.csv")
Chem.Phys <- read.csv("./Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv")

# Set date to date format
#EPAair$Date <- as.Date(EPAair$Date, format = "%Y-%m-%d")

#2
RMtheme <- theme_bw(base_size = 11) +
  theme(plot.title = element_text(size = 15, color = "black", hjust = .5),
    axis.text = element_text(color = "black"),
    legend.position = "bottom", legend.text = element_text(size = 9), legend.title = element_text(size =
theme_set(RMtheme)
```

Neonicotinoids test

Research question: Were studies on various neonicotinoid chemicals conducted in different years?

3. Generate a line of code to determine how many different chemicals are listed in the Chemical.Name column.
4. Are the publication years associated with each chemical well-approximated by a normal distribution? Run the appropriate test and also generate a frequency polygon to illustrate the distribution of counts for each year, divided by chemical name. Bonus points if you can generate the results of your test from a pipe function. No need to make this graph pretty.
5. Is there equal variance among the publication years for each chemical? Hint: var.test is not the correct function.

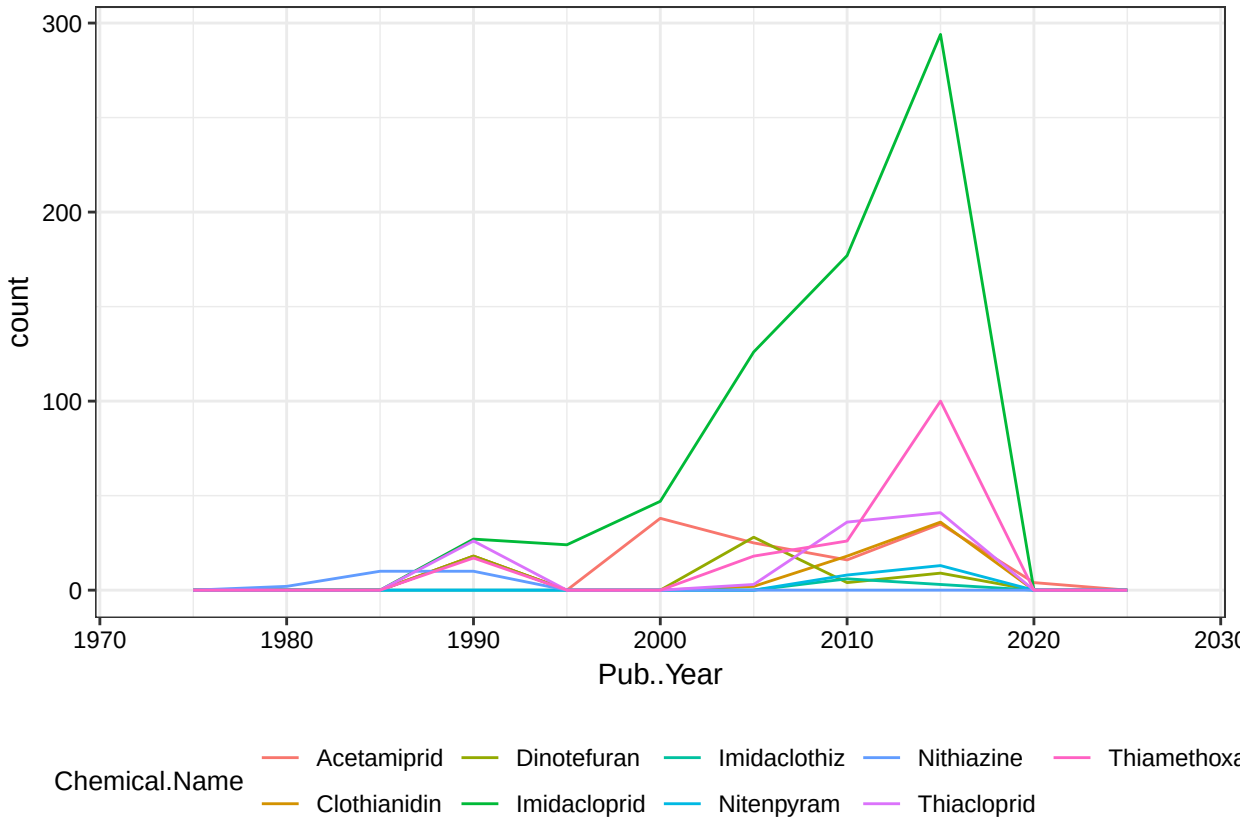
```r
#3
summary(EPAecotox$Chemical.Name)
```

```
##  Acetamiprid Clothianidin  Dinotefuran Imidacloprid Imidaclothiz
##          136           74           59          695            9
##   Nitenpyram   Nithiazine  Thiacloprid Thiamethoxam
##           21           22          106          161
```

```r
#4
shapiro.test(EPAecotox$Pub..Year)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  EPAecotox$Pub..Year
## W = 0.85472, p-value < 2.2e-16
```

```r
ggplot(EPAecotox, aes(x = Pub..Year, color = Chemical.Name)) +
  geom_freqpoly(binwidth = 5)
```




```
#5
bartlett.test(EPAecotox$Pub..Year,EPAecotox$Chemical.Name)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  EPAecotox$Pub..Year and EPAecotox$Chemical.Name
## Bartlett's K-squared = 139.59, df = 8, p-value < 2.2e-16
```

```
# OR bartlett.test(Pub..Year~Chemical.Name,EPAecotox) works too
```

6. Based on your results, which test would you choose to run to answer your research question?

   ANSWER: The shapiro wilk test with a p-value of less than .001 and the corresponding frequency polygon results tell me that the years that studies were published are not normally dristributed but rather are skewed. The bartlett test result has a p-value of less than .001 as well. This means I can reject the null hypothesis that the variance is the same for all chemicals, and there is evidence to suggest that the variance in publication years is different for all the chemicals. Since the data is not normally distributed and the variances are not equal, I will use the non-parametric equivalent of ANOVA: the Kruskal-Wallis Test.

7. Run this test below.

8. Generate a boxplot representing the range of publication years for each chemical. Adjust your graph to make it pretty.

```
#7
Pub..Year.kw <- kruskal.test(EPAecotox$Pub..Year ~ EPAecotox$Chemical.Name)
Pub..Year.kw
```

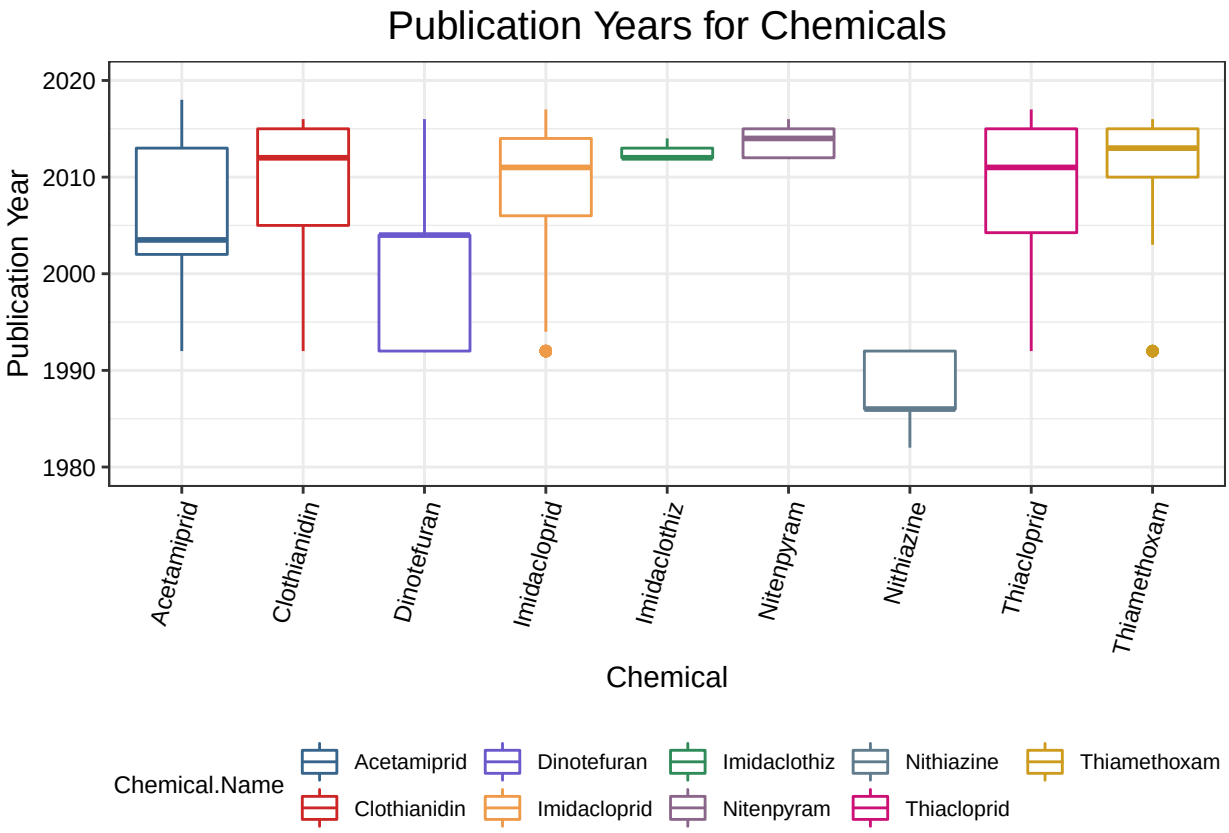
```
##
##  Kruskal-Wallis rank sum test
##
## data:  EPAecotox$Pub..Year by EPAecotox$Chemical.Name
## Kruskal-Wallis chi-squared = 134.15, df = 8, p-value < 2.2e-16
```

```
dunn.test(EPAecotox$Pub..Year, EPAecotox$Chemical.Name, kw = T, table = F, list = T, method = "holm", al
```

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 134.1455, df = 8, p-value = 0
##
##
##                            Comparison of x by group
##                                    (Holm)
##
## List of pairwise comparisons: Z statistic (adjusted p-value)
## ------------------------------------------------
## Acetamiprid - Clothianidin  : -3.038807 (0.0404)*
## Acetamiprid - Dinotefuran   :  2.117208 (0.4109)
## Clothianidin - Dinotefuran  :  4.406076 (0.0002)*
## Acetamiprid - Imidacloprid  : -4.020498 (0.0013)*
## Clothianidin - Imidacloprid :  0.506889 (1.0000)
## Dinotefuran - Imidacloprid  : -5.214028 (0.0000)*
## Acetamiprid - Imidaclothiz  : -1.805293 (0.7813)
## Clothianidin - Imidaclothiz : -0.516664 (1.0000)
## Dinotefuran - Imidaclothiz  : -2.658649 (0.1177)
## Imidacloprid - Imidaclothiz : -0.728428 (1.0000)
## Acetamiprid - Nitenpyram    : -4.501863 (0.0002)*
## Clothianidin - Nitenpyram   : -2.493626 (0.1770)
## Dinotefuran - Nitenpyram    : -5.452779 (0.0000)*
## Imidacloprid - Nitenpyram   : -3.063483 (0.0394)*
## Imidaclothiz - Nitenpyram   : -1.089720 (1.0000)
## Acetamiprid - Nithiazine    :  5.642529 (0.0000)*
## Clothianidin - Nithiazine   :  7.147325 (0.0000)*
## Dinotefuran - Nithiazine    :  3.869350 (0.0023)*
## Imidacloprid - Nithiazine   :  7.728634 (0.0000)*
## Imidaclothiz - Nithiazine   :  4.847313 (0.0000)*
## Nitenpyram - Nithiazine     :  7.709981 (0.0000)*
## Acetamiprid - Thiacloprid   : -3.222561 (0.0241)*
## Clothianidin - Thiacloprid  :  0.141491 (0.8875)
## Dinotefuran - Thiacloprid   : -4.602529 (0.0001)*
## Imidacloprid - Thiacloprid  : -0.388871 (1.0000)
## Imidaclothiz - Thiacloprid  :  0.587068 (1.0000)
## Nitenpyram - Thiacloprid    :  2.670974 (0.1210)
## Nithiazine - Thiacloprid    : -7.316688 (0.0000)*
## Acetamiprid - Thiamethoxam  : -5.889886 (0.0000)*
## Clothianidin - Thiamethoxam : -1.758725 (0.7862)
## Dinotefuran - Thiamethoxam  : -6.676212 (0.0000)*
## Imidacloprid - Thiamethoxam : -3.532703 (0.0082)*
## Imidaclothiz - Thiamethoxam : -0.188627 (1.0000)
## Nitenpyram - Thiamethoxam   :  1.592776 (1.0000)
## Nithiazine - Thiamethoxam   : -8.722412 (0.0000)*
## Thiacloprid - Thiamethoxam  : -2.146115 (0.4142)
```

```
##
## alpha = 0.05
## Reject Ho if p <= alpha
```

```
#8
legend_title_7 <- "Chemical Name"
Chem.Pub.Box <-
  ggplot(EPAecotox, aes(x = Chemical.Name, y = Pub..Year, color = Chemical.Name)) +
  geom_boxplot() +
  scale_color_manual(values = c("steelblue4", "firebrick3", "slateblue3", "tan2", "seagreen4", "plum4",
  scale_y_continuous(limits = c(1980,2020)) +
  ggtitle("Publication Years for Chemicals") +
  ylab(expression("Publication Year")) +
  xlab(expression("Chemical")) +
  theme(axis.text.x = element_text(angle = 75, hjust = 1), legend.text = element_text(size = 8), legend
print(Chem.Pub.Box)
```



9. How would you summarize the conclusion of your analysis? Include a sentence summarizing your findings and include the results of your test in parentheses at the end of the sentence.

   ANSWER: Yes, studies on various neonicotinoid chemicals were conducted in different years. The result of the Kruskal-Wallis test indicates that there is a significant differene between groups (chemicals), though the test does not indicate which chemicals differ from each other (Df = 8, p-value < .001 or 2.2e-16). Looking at the box plot you can see that Nithiazine studies occurred much earlier than other studies. Dinotefuran studies were on the earlier end as were Acetamiprid studiest, but many of the other chemicals had studies that took place around or shortly after 2010. The studies were not normally distributed, per the result of the Shapiro-Wilk test with a p-value of less than .001. The results of the Bartlett test also had a p-value of less than .001, meaning

the variances of chemicals are not equal. This was confirmed with the Dunn test which did a pair-wise comparison of variances. As depcited in the box plot, the pairs involving Nithiazine all had p-values of less than .00001, meaning they did no have equal variances, evan as some pairs showed to have similar variances with p-values close to 1.

NTL-LTER test

Research question: What is the best set of predictors for lake temperatures in July across the monitoring period at the North Temperate Lakes LTER?

11. Wrangle your NTL-LTER dataset with a pipe function so that it contains only the following criteria:

- Only dates in July (hint: use the daynum column). No need to consider leap years.
- Only the columns: lakename, year4, daynum, depth, temperature_C
- Only complete cases (i.e., remove NAs)

12. Run an AIC to determine what set of explanatory variables (year4, daynum, depth) is best suited to predict temperature. Run a multiple regression on the recommended set of variables.

```
#11
Chem.Phys.Filter <-
  Chem.Phys %>%
  filter(daynum %in% c(183:212)) %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  filter(!is.na(lakename) & !is.na(year4) &!is.na(daynum) & !is.na(depth) & !is.na(temperature_C))

#12
Chem.Phys_AIC <- lm(data = Chem.Phys.Filter, temperature_C ~ year4 + daynum + depth)
step(Chem.Phys_AIC)
```

```
## Start:  AIC=25233.58
## temperature_C ~ year4 + daynum + depth
##
##          Df Sum of Sq    RSS   AIC
## <none>                137124 25234
## - year4   1       115 137239 25239
## - daynum  1      1015 138139 25301
## - depth   1    392438 529563 37958
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = Chem.Phys.Filter)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##   -10.13919      0.01232      0.03789     -1.94770
```

```
Chem.Phys_model <- lm(data = Chem.Phys.Filter, temperature_C ~ year4 + daynum + depth)
summary(Chem.Phys_model)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = Chem.Phys.Filter)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6680 -3.0016  0.0914  2.9773 13.6150
##
```

```
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -10.139191   8.801260   -1.152  0.24934
## year4         0.012323   0.004385    2.810  0.00496 **
## daynum        0.037893   0.004539    8.348  < 2e-16 ***
## depth        -1.947704   0.011865 -164.149  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.816 on 9415 degrees of freedom
## Multiple R-squared:  0.7416, Adjusted R-squared:  0.7415
## F-statistic:  9008 on 3 and 9415 DF,  p-value: < 2.2e-16
```

13. What is the final linear equation to predict temperature from your multiple regression? How much of the observed variance does this model explain?

    ANSWER: Given that AIC was lowest when no variables were removed, I used the linear equation with all three varibles to predict temperature: Temperature_C = -10.139 + .012(year4) + .038(daynum) - 1.948(depth). This model explains 74.15% of the variance in temperature.

14. Run an interaction effects ANCOVA to predict temperature based on depth and lakename from the same wrangled dataset.

```
#14
Depth_Lake.interaction <- lm(data = Chem.Phys.Filter, temperature_C ~ lakename * depth)
summary(Depth_Lake.interaction)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename * depth, data = Chem.Phys.Filter)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6773 -2.8928 -0.2863  2.7567 16.3606
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      22.9455     0.5860  39.156  < 2e-16 ***
## lakenameCrampton Lake             2.2173     0.6802   3.260 0.001119 **
## lakenameEast Long Lake           -4.3884     0.6189  -7.090 1.43e-12 ***
## lakenameHummingbird Lake         -2.8915     0.8634  -3.349 0.000814 ***
## lakenamePaul Lake                 0.6607     0.5986   1.104 0.269716
## lakenamePeter Lake                0.3459     0.5973   0.579 0.562564
## lakenameTuesday Lake             -2.8622     0.6066  -4.718 2.41e-06 ***
## lakenameWard Lake                 2.4180     0.8432   2.868 0.004145 **
## lakenameWest Long Lake           -2.3753     0.6184  -3.841 0.000123 ***
## depth                            -2.5820     0.2410 -10.713  < 2e-16 ***
## lakenameCrampton Lake:depth       0.8058     0.2465   3.269 0.001083 **
## lakenameEast Long Lake:depth      0.9465     0.2432   3.892 0.000100 ***
## lakenameHummingbird Lake:depth   -0.4840     0.2971  -1.629 0.103394
## lakenamePaul Lake:depth           0.4005     0.2421   1.655 0.098027 .
## lakenamePeter Lake:depth          0.5792     0.2418   2.395 0.016619 *
## lakenameTuesday Lake:depth        0.6574     0.2426   2.710 0.006737 **
## lakenameWard Lake:depth          -0.6930     0.2861  -2.422 0.015457 *
## lakenameWest Long Lake:depth      0.8090     0.2432   3.327 0.000883 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.47 on 9401 degrees of freedom
## Multiple R-squared:  0.7867, Adjusted R-squared:  0.7863
## F-statistic:  2040 on 17 and 9401 DF,  p-value: < 2.2e-16
```

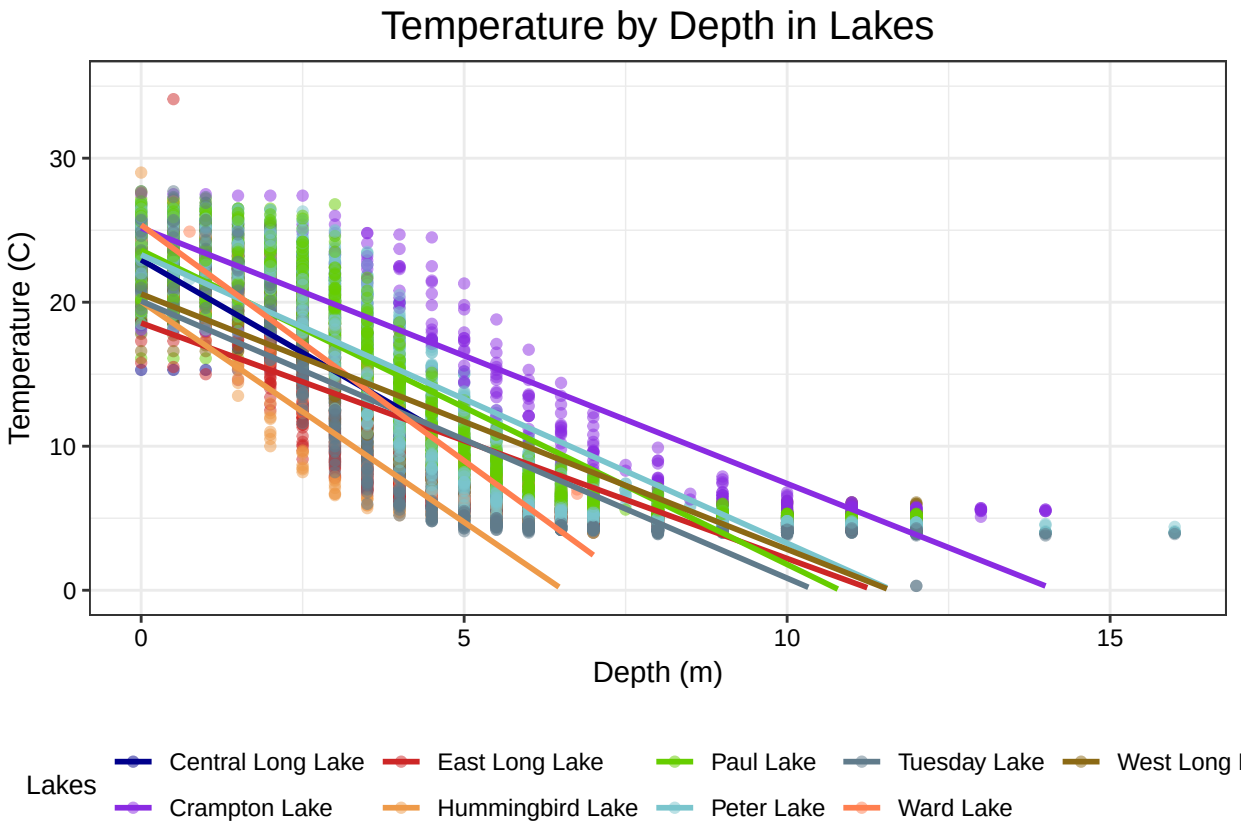
15. Is there an interaction between depth and lakename? How much variance in the temperature observations does this explain?

    ANSWER: Yes, there is an interaction between depth and lakename. The p-value for the interaction effect is less than 0.05, thus we consider the interaction between depth and lakename to be significant. The interaction between lakename and depth explains 78.63% of the variance in temperature.

16. Create a graph that depicts temperature by depth, with a separate color for each lake. Add a geom_smooth (method = "lm", se = FALSE) for each lake. Make your points 50 % transparent. Adjust your y axis limits to go from 0 to 35 degrees. Clean up your graph to make it pretty.

```
#16
legend_title_TxD <- "Lakes"
TempXDepth.graph <-
  ggplot(Chem.Phys.Filter, aes(x = depth, y = temperature_C, color = lakename)) +
  geom_point(alpha = .5) +
  ggtitle("Temperature by Depth in Lakes") +
  ylab(expression("Temperature (C)")) +
  scale_y_continuous(limits = c(0,35)) +
  xlab(expression("Depth (m)")) +
  scale_x_continuous(breaks=seq(0, 25, by=5)) +
  scale_color_manual(legend_title_TxD,values = c("darkblue","blueviolet","firebrick3", "tan2", "chartreu
  geom_smooth(aes(color = lakename),method = lm, se= FALSE)
print(TempXDepth.graph)
```

```
## Warning: Removed 72 rows containing missing values (geom_smooth).
```


Temperature by Depth in Lakes
Temperature (C)
30
20
10
0
0
5
10
15
Depth (m)
Lakes
Central Long Lake
East Long Lake
Paul Lake
Tuesday Lake
West Long
Crampton Lake
Hummingbird Lake
Peter Lake
Ward Lake